DOCUMENT RESUME

ED 411 264 TM 027 212

AUTHOR Lambert, Richard; Flowers, Claudia; Sipe, Theresa; Idleman, Lynda

TITLE Integrating Software for Sample Size Calculations, Data Entry, and Tabulation: Software Demonstration of a System for Survey Research.

PUB DATE 1997-03-00

NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Computation; *Computer Software; Data Analysis; *Integrated Activities; *Research Methodology; *Sample Size; *Surveys

IDENTIFIERS *Epi Info Computer Program; *Statistical Package for the Social Sciences

ABSTRACT

This paper discusses three software packages that offer unique features and options that greatly simplify the research package for conducting surveys. The first package, EPSILON, from Resource Group, Ltd. of Dallas (Texas) is designed to perform a variety of sample size calculations covering most of the commonly encountered survey research situations. The second package, Epi Info, is a word-processing database and statistics program created by the Centers for Disease Control and Prevention (CDC). This program is shareware and is available from the CDC homepage. The third package, the Statistical Package for the Social Sciences (SPSS), is an integrated system for managing, analyzing, and displaying data. This paper describes how to import Epi Info generated files into major spreadsheet packages and statistical packages, including some specific pointers on how to modify Epi Into generated SPSS programming code for optimal use in SPSS for Windows. Also discussed is using EPSILON software for sampling related calculations. A sample case is presented for a survey of a professional organization with 7,500 members. The steps for entering the data into Epi Info and then retrieving the Epi Info information into SPSS are outlined. (Contains two tables.) (SLD)

Integrating Software for Sample Size Calculations, Data Entry, and Tabulation:

Software Demonstration of a System for Survey Research

Richard Lambert
Claudia Flowers
University of North Carolina Charlotte

Theresa Sipe
Georgia State University

Lynda Idleman
Idleman and Associates

Paper presented at the 1997 Annual Convention of the
American Educational Research Association
Chicago, IL

Correspondence concerning this paper should be directed to:
Richard G. Lambert, Ph.D.
Department of Educational Administration, Research, and Technology
University of North Carolina at Charlotte
Charlotte, NC 28223-0001
rglamber@email.uncc.edu

Integrating Software for Sample Size Calculations, Data Entry, and Tabulation: Software Demonstration of a System for Survey Research

Recent years have shown a large proliferation of software products to aid the survey researcher. However, few individual products can offer comprehensive assistance from sample size determinations, survey design, data collection and interviewing, data entry, and tabulation. This presentation demonstrates three software packages which offer unique features and options that greatly simplify the survey research process. While not constituting a single software package, the tools to be demonstrated form a unified system for survey research.

The first package, EPSILON, from Resource Group, Ltd. of Dallas, Texas is designed to perform a variety of sample size calculations covering most of the commonly encountered survey research situations and is very useful for designing sampling plans. Copies made be obtained from the following address:

Resource Group, Ltd.
Three Lincoln Centre
Suite 1600
5430 LBJ Freeway
Dallas, TX 75240
phone: 800-221-6308
fax: 800-238-8849

The second software package, Epi Info, is a word-processing database, and statistics program created by the Centers for Disease Control and Prevention (CDC). It is shareware and can be downloaded from the CDC homepage (www.cdc.gov). The manual and diskettes of the software can be purchased for nominal cost from USD Incorporated (770-469-4098). The manual is contained within the software. The CDC also provides a Hotline for Technical Assistance (404-728-0545). Epi Info includes a word processor for creating questionnaires. It also contains a data entry module that can be linked to the word processor and can be used to create data entry screens that can appear exactly as the paper copies of the questionnaire. This makes data entry less error prone and can also be used as a simple CATI or on-screen interviewing technique in which completed interviews are added to the data set as they are completed. It allows for double entry and verification as well as coding, skip patterns, and open ended questions. Epi Info also includes a series of tabulation, statistical analysis, and graphing modules including routines for random number generation and exact confidence intervals. It has 12 different file formats for export and actually writes part of the SAS or SPSS code for the user if these file formats are selected.

The third software package, SPSS, is a comprehensive, integrated system for managing, analyzing, and displaying data. For more information call 312-329-3500. Other commerce statistical software packages (e.g., SAS) could easily be substituted into this demonstration but due to a time limit, only SPSS will be demonstrated. This paper describes how to import Epi Info

generated files into major spreadsheet packages and statistical packages, including some specific pointers on how to modify Epi Info generated SPSS programming code for optimal use in SPSS for Windows.

Epsilon Software for Sampling Related Calculations

Features

The program Epsilon allows you to perform various sampling related calculations quickly and easily. If you need to know how many subjects to include in an upcoming study, the program can help provide some guidelines. If you need to report the actual precision you obtained in a study already completed, this too is available. You can calculate sample size needed, precision obtained, and standard error for both Simple Random Sampling and Stratified Random Sampling. Weighting, Coefficient of Variation, z-tests, and an assortment of tools are also offered. The most frequently requested sampling formulae are conveniently at your disposal, requiring minimum effort and offering you the opportunity to perform various simulations and "what if" scenarios.

All program options that calculate n, or desired sample size, can be used for estimating means or proportions. The following functions appear across a variety of situations. Here are brief explanations of what they perform:

1.) Calculate n (sample size needed) for a given epsilon, tolerable error between obtained estimates and the population parameter. This indicates how many subjects are required in order to obtain a given level of precision.

2.) Calculate epsilon (tolerable error) for a given n (sample size). These functions do the inverse of that above and enable you to do some simulation determining what level of precision would be expected if you were to obtain a certain number of subjects. They also allow you to report the actual precision obtained after a study has been conducted. You may be in a better position to input a standard deviation estimate after the data has been collected.

3.) Calculate Standard Error of the Mean with the Finite Population Correction (FPC) factor. Many statistical packages do not provide standard error calculations using the FPC. The standard error of the mean is impacted by the sampling method, while most statistical packages assume Simple Random Sampling from an infinite population or Simple Random Sampling with replacement. These options can be used to determine Standard Errors using the FPC and considering the sampling method. These options also provide a confidence interval (CI) around the mean that is based on the standard error figured with the FPC and sampling method considerations. The z value you enter corresponds to the desired CI.

The program contains four main sections: 1.) Simple Random Sampling, 2.) Stratified

Random Sampling, 3.) Tools and Tests, and 4.) Random Number Generators. Each section is described in detail below.

Simple Random Sampling

To use this section you must first determine if it is necessary to include the FPC in your calculations. As discussed above, if you are sampling without replacement and plan to sample more than 5 % of the population, the FPC options will be helpful. The options on the top half of the screen employ the FPC, those on the bottom half of the menu do not. Next you need to determine if you are interested in estimating confidence intervals around single parameter estimates (populations means or proportions), or around the differences between pairs of parameter estimates. The options on the right side of the screen are to be used when interest lies in estimating the actual magnitude of a difference between population means or proportions. The left side of the screen is used when considering a confidence interval around a single parameter estimate. You will be prompted to enter your estimate of the population standard deviation, epsilon, and values for N, n, range, and desired % error as needed. The program will calculate epsilon based upon your entry of desired percent error and range. Specific features are as follows:

a. Sample size needed to obtain a given sampling error, with or without the finite population correction factor, when interested in estimating a single mean or proportion.

b. Sampling error obtained for a given sample size, with or without the finite population correction factor, when interested in estimating a single mean or proportion.

c. Standard Error and Confidence Interval around a mean or proportion, with or without the finite population correction factor.

d. Sample size needed to obtain a given sampling error, with or without the finite population correction factor, when interested in estimating the difference between two means or two proportions.

e. Sampling error obtained for a given sample size, with or without the finite population correction factor, when interested in estimating the difference between two means or two proportions.

f. Standard Error and Confidence Interval around the difference between two means or two proportions, with or without the finite population correction factor.

Stratified Random Sampling

This section offers four different ways to calculate n. They differ in their goals and allocation schemes. Allocation is the process of determining how many respondents to sample within each strata. All of these options use the FPC and require you to enter an actual epsilon value, not a range and percent error. Again, epsilon refers to the amount of error you are willing to tolerate in either direction around your estimate. It must be measured in the same units as the mean in question. Each option uses the symbol K to indicate the number of strata. This section will calculate epsilon for a given n and standard error, just as in the Simple Random Sampling section. The four allocation methods are as follows:

1.) Proportional Allocation seeks to minimize overall n for a given epsilon and allocates to strata based on the proportions of the population that each stratum contains.

2.) Optimum Allocation - Seeks to minimize overall n for a given epsilon while allowing allocation to strata to become whatever is necessary to minimize n. This method can adjust for differences in variance between strata.

3.) Neyman Allocation - Optimum allocation is a special case of Neyman Allocation when the cost of obtaining respondents in each stratum is equal. This method allows for consideration of the cost of obtaining respondents in each stratum. It seeks to minimize overall n while minimizing cost.

4.) Specified Allocation - Offers the general form of the formula for overall n given stratified random sampling. It allows the researcher to specify the allocation scheme and then minimizes overall n while meeting the requested error and allocation scheme.

Tools and Tests

This section contains a variety of z test options including one mean, two means, one proportion, and two proportions. Each will display one tailed and two probabilities as well as allow the user to specify the hypothesized amount of difference to test. These features are useful when reading through a set of standard market research crosstabs. While many of the tabulation packages allow for significance testing to be placed on the crosstab, the reports may not always include them. The z test functions can be used to get a rough guess as to whether differences are beyond what would expected due to sampling error alone in order to determine if significance testing is warranted. Z tests are in a sense liberal relative to t tests in that with small sample sizes the critical t will exceed the critical z. Therefore if a difference is not big enough to achieve significance with a z test it will not do so with a t test either. If a difference is big enough to achieve significance with a z test it will not necessarily active significance with a t test. The tests serve as general guideline that can be quickly calculated without further programming. Also included in this section are factorials, combinations, permutations, weighting, and proportions of the total area under sections of the normal curve.

Random Number Generators

The program will generate random numbers from a uniform distribution in a variety of forms including values between 0 and 1, random starting point in a list of any length, and random digits up to 7 digits in length. It will also generate phone number extensions to an exchange that you provide and will delete duplicate numbers as it generates. Random values generated from a standard normal distribution are also available along with corresponding transformations to any population mean and standard deviation. Each of these options can be saved to a file, sent to the screen, or send to a printer.

A Sample Case

If you are estimating parameters for several items on a questionnaire, as is almost always the case, the question of how to calculate sample size is more complex than if you were interested in a single parameter estimate. One standard suggestion is to use the values related to the most important question. Other strategies include using a maximum estimate and using the largest value across the items. Obtaining an estimate of standard deviation from pilot study data, previous research, or published statistics can be very useful. Of course this is not always possible, especially with questionnaires that have never been used before. The method of choosing the maximum estimate is common and conservative, although maximum variance rarely actually occurs in real data. Some have suggested that an easy method is to divide the range of possible answers to a question by 6, since six standard deviation units comprise almost all of the area in a normal distribution. If you choose to use the maximum estimate method and you are estimating a proportion, remember that the maximum population standard deviation of a dichotomous (0/1) variable, is .5. The maximum population standard deviation of a question for which means will be calculated is the range, that is the highest possible value minus the lowest value, divided by 2.

Consider the case of conducting a survey for a professional organization of 7,500 member institutions. Suppose a pilot study was conducted, the most important question identified, and means and standard deviations calculated from pilot data. Let us also assume that the organization is interested in stratified sampling and has identified two stratification variables of interest, membership status and organization type. Table 1 illustrates the pilot study results along with the percentage of the population contained in each strata. Table 2 shows the sampling plan that would be provided by the software. If the proportional allocation option with a finite population was selected, the program would prompt the user for the number of strata, 4, the respective strata weights or proportions of the population they represent, and the strata standard deviations. In this case the pilot data would be supplied.

Insert Tables 1 and 2 about here

Epi Info Data Management Program

The following are steps in entering data into Epi Info. The tutorial is excellent in teaching how to use the software.

1. Create a ".QES" file in the editor. This file provides the structure for the data file (i.e., the ".qes" file describes the data to be entered). From the main menu choose EPED to enter the editor. For each variable type in the variable name and indicate the number of columns needed for the variable. A "#" is used for numeric data and a "_" is used for alphanumeric data. The saved file name must have the ".qes" extension.

2. Create a ".REC" file. The ".rec" file is where the data are entered. From the main menu choose ENTER. Type in the name of your file you wish to create. Choose "2" to create a file. Type in the name of the ".qes" file that will provide the structure for the data file. A ".rec" file will be created. You are ready to enter data.

3. To set limits on certain fields choose CHECK from the main menu. Type in the name of the ".rec" file. The codes are at the bottom of the screen. To set a minimum and maximum value for a certain variable go to that variable field, type the minimum value possible (i.e., 1) press F1 key. Next type the maximum value possible (i.e., 4) and press the F2 key. See the tutorial or manual for other options available in CHECK.

4. To do double data entry for validation, create two ".rec" files (i.e., data1.rec and data2.rec). Both can be created from the same ".qes" file. Enter the data in both files. Choose VALIDATE from the main menu. Enter the name of file 1 and file 2, choose where the results will be displayed, and choose the name of the linking variable (usually the ID number). The files will be compared and the results will indicate where there are discrepancies.

5. To export the files into another format choose EXPORT from the main menu. Type in the name of the ".rec" file and choose the type of file you wish (i.e., SPSS-PC, SAS, dBase, etc.). Epi Info will create the new file (i.e., data1.sps for a SPSS file, data1.sas for a SAS file). The rec file will be preserved. EXPORT will allow you to make as many new files as you wish.

6. Epi Info will also do simple data analysis from the ANALYSIS option. See the tutorial and manual for the options.

The CDC encourages this program to be shared so give it to your friends and colleagues.

Retrieving Epi Info Files into SPSS

Epi Info will export data to a variety of file formats including SAS, SPSS, Lotus, and dBase. In order to export to another package simply select the export option, select the file type, identify the *.REC file to be converted, and identify the destination directory and file name for the exported file. When selecting an SPSS file format, there are a few issues to be aware of. The file that is exported is actually a program that can be run to create an SPSS dataset. The program contains a DATA LIST statement along with the data embedded within a BEGIN DATA statement. The syntax needs to be modified in a few ways in order to actually run within SPSS for Windows. The example below illustrates the changes (in bold) to be made.

*.SPS Program File as written by EPI Info:

```
DATA LIST
   /ID 1-3 Q1 4 Q2 6
   /Q3 1 Q4 3.
BEGIN DATA
101 2 3
2 3
102 3 4
3 2
103 2 4
3 4
END DATA.
```

*.SPS Program File as Modified to Run in SPSS for Windows:

```
DATA LIST RECORDS=2
   /1 ID 1-3 Q1 4 Q2 6
   /2 Q3 1 Q4 3.
BEGIN DATA
101 2 3
2 3
102 3 4
3 2
103 2 4
3 4
END DATA.
SAVE OUTFILE = "C:\ * \ *.SAV".
```

The changes involve adding the RECORDS=(n) statement and affixing the record number after the / within each line of the DATA LIST statement. The SAVE OUTFILE statement will then save a *.SAV file that can be retrieved with subsequent programs.

Table 1.
Pilot Study Results for the Most Important Question.

		Institution Type	
Membership Status	Statistic	Academic	Other
Full	Pilot Mean	3.57	3.42
	Pilot SD	1.05	1.67
	Pop. %	47.52%	16.84%
Associate	Pilot Mean	3.47	3.22
	Pilot SD	1.08	1.59
	Pop. %	22.06%	13.58%

Table 2.
Sampling Plan.

		Institution Type	
Membership Status	Statistic	Academic	Other
Full	Desired n	192	68
	Exp. Resp. %	70.00%	50.00%
	Sent Out	274	136
Associate	Desired n	89	55
	Exp. Resp. %	70.00%	50.00%
	Sent Out	127	110

Note. - Desired sample sizes are based upon 95% CI with +/- 3% error and a population size of 7,500.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

TM027212

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: INTEGRATING SOFTWARE FOR SAMPLE SIZE CALCULATIONS, DATA ENTRY, AND TABULATION: SOFTWARE DEMONSTRATION OF A SYSTEM FOR SURVEY RESEARCH

Author(s): LAMBERT, R., FLOWERS, C., SIPE, T., AND IDLEMAN, L.

Corporate Source: 1997 AERA CONVENTION

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

☐ Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature:

Printed Name/Position/Title:
RICHARD G. LAMBERT, PH.D.
ASSISTANT PROFESSOR

Organization/Address:
DEPT. OF ED. ADMIN., RESEARCH, AND TECHNOLOGY
UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE
CHARLOTTE, NC 28223-0001

Telephone: 704-547-3735

FAX: 704-5

E-Mail Address: RGLAMBERT EMAIL.UNCC.EDU

Date: 4/16/97